Estimating Smooth GLM in Non-interactive Local Differential Privacy Model with Public Unlabeled Data

Di Wang * SUNY at Buffalo
Huanyu Zhang Cornell University
Marco Gaboardi Boston University
Jinhui Xu SUNY at Buffalo

Abstract

In this paper, we study the problem of estimating smooth Generalized Linear Models (GLM) in the Non-interactive Local Differential Privacy (NLDP) model. Different from its classical setting, our model allows the server to process some additional public but unlabeled data. By using Stein's lemma and its variants, we first show that there is an (ϵ, δ)-NLDP algorithm for GLM (under some mild assumptions), if each data record is i.i.d sampled from some sub-Gaussian distribution with bounded ℓ_1-norm. Then with high probability, the sample complexity of public and private data, for the algorithm to achieve an α estimation error (in ℓ_∞-norm), is $O(p^2\alpha^{-2})$ and $O(p^2\alpha^{-2}\epsilon^{-2})$ respectively if α is not too small (*i.e.,* $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$), where p is the dimensionality of the data. This is a significant improvement over the previously known quasi-polynomial (in α) or exponential (in p) complexity of GLM with no public data. We demonstrate the effectiveness of our algorithms through experiments on both synthetic and real world datasets.

1 Introduction

Generalized Linear Model (GLM) is one of the most fundamental models in statistics and machine learning. It generalizes ordinary linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. GLM was introduced as a way of unifying various statistical models, including linear, logistic and Poisson regressions.

GLM: Let $y \in [0, 1]$ be the response variable that belongs to an exponential family with natural parameter η. That is, its probability density function could be written as $p(y|\eta) = \exp(\eta y - \Phi(\eta))h(y)$, where Φ is the *cumulative generating function*. Given observations $y_1, \cdots, y_n$ such that $y_i \sim p(y_i|\eta_i)$ for $\eta = (\eta_1, \cdots, \eta_n)$, the maximum likelihood estimator (MLE) can be written as $p(y_1, y_2, \cdots|\eta) = \exp(\sum_{i=1}^n y_i\eta_i - \Phi(\eta_i))\Pi_{i=1}^n h(y_i)$. In GLM, we assume that η is modeled by linear relations, *i.e.,* $\eta_i = \langle x_i, w^*\rangle$ for some $w^* \in \mathbb{R}^p$ and feature vector x_i. Thus, maximizing MLE is equivalent to minimizing $\frac{1}{n}\sum_{i=1}^n[\Phi(\langle x_i, w\rangle) - y_i\langle x_i, w\rangle]$. The goal is to find w^*, which is equivalent to minimizing its population version

$$w^* = \arg\min_{w\in\mathbb{R}^p} \mathbb{E}_{(x,y)}[\Phi(\langle x, w\rangle) - y\langle x, w\rangle]. \tag{1}$$

One often encountered challenge for using GLM in real world applications is how to handle sensitive data, such as those in social science and medical research. As a commonly-accepted approach for preserving privacy, Differential Privacy (DP) [9] provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers.

As a popular way of achieving DP, Local Differential Privacy (LDP) has received considerable attentions in recent years and been adopted in industry [7, 11, 20]. In LDP, each individual manages her proper data and discloses them to a server through some DP mechanisms. The server collects the

*The first two authors contribute equally. Part of the work was done when D.W and M.G were visiting Simons Institute for the Theory of Computing.

Preprint. Under review.

(now private) data of each individual and combines them into a resulting data analysis. Information exchange between the server and each individual could be either only once or multiple times. Correspondingly, protocols for LDP are called non-interactive LDP (NLDP) or interactive LDP. Due to its ease of implementation (*e.g.* no need to deal with the network latency), NLDP is often preferred in practice.

While there are many results on GLM in the DP and interactive LDP models [5, 1, 13, 14], GLM in NLDP is still not well understood due to the limitation of the model. [18, 23, 26] and [24] comprehensively studied this problem. However, all of the results are on the negative side. More specifically, they showed that to achieve an error of α, the sample complexity needs to be quasi-polynomial in α [24, 26] or even exponential in the dimensionality p [18, 23] (see Related Work section for more details). Due to these negative results, there is no study on the practical performance on these algorithms.

On the other hand, instead of the classical DP model or its relaxations, some recent work focus on a relaxed model of DP where the server has additional public but unlabeled data, such as [**?** 12, 16, 17, 2]. Specifically, they show that with the power of these public unlabeled data, the sample complexity could be further improved [2] under the assumption that these public data has the same marginal distribution as the private ones. And it has better practical performance than the classical DP model on some tasks, such as the Empirical Risk Minimization [12, 16]. However, all of above work focus on the central DP model, thus there is no existing work study the NLDP model with public unlabeled data.

Thus, a natural question is: **For the problem of estimating GLM in the NLDP model, can we further reduce the sample complexity if the curator has additional public but unlabeled data? Moreover, is there any efficient and effective algorithm on the problem?**

In this paper, we provide a partial answer to the above two questions by studying the NLDP model where the curator (server) is allowed to access some additional *public but unlabeled data*. Our contributions can be summarized as follows:

1. We first show that when the feature vector x of GLM is sub-Gaussian with bounded ℓ_1-norm, there is an (ϵ, δ)-NLDP algorithm for GLM (under some mild assumptions) whose sample complexity for achieving an error of α (in ℓ_∞-norm) is $O(p^2\epsilon^{-2}\alpha^{-2})$ and $O(p^2\alpha^{-2})$ (with other terms omitted) for private and public data respectively, if α is not too small (*i.e.,* $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$). We note that this is the first result that achieves a **fully polynomial** sample complexity for a general class of loss functions in the NLDP model with public unlabeled data.
2. Then we provide an experimental study of our algorithm on both synthetic and real world datasets. The experimental results suggest that our methods are efficient and effective, which is consistent with our theoretical analysis. To our best knowledge, these are the **first** effective algorithms in the NLDP model with public unlabeled data for GLM problem.

2 Non-interactive LDP Model with Public Unlabeled Data

Local Differential Privacy (LDP): In LDP, we have data universe $\mathcal{X}$ and $\mathcal{Y}$, n players with each holding a private data record $(x, y) \in \mathcal{X} \times \mathcal{Y}$ sampled from some distribution $\mathcal{P}$, where $x \in \mathbb{R}^p$ is the feature vector and $y \in \mathbb{R}$ is the label or response, and a server that is in charge of coordinating the protocol. An LDP protocol proceeds in T rounds. In each round, the server sends a message, which is often called a query, to a subset of the players, requesting them to run a particular algorithm. Based on the query, each player i in the subset selects an algorithm Q_i, runs it on her own data, and sends the output back to the server.

Definition 1. [15] An algorithm Q is ϵ-locally differentially private (LDP) if for all pairs $x, x' \in \mathcal{D}$, and for all events E in the output space of Q, we have $\Pr[Q(x) \in E] \leq e^\epsilon \Pr[Q(x') \in E]$. A multi-player protocol is ϵ-LDP if for all possible inputs and runs of the protocol, the transcript of player i's interaction with the server is ϵ-LDP. If $T = 1$, we say that the protocol is ϵ **non-interactive LDP (NLDP)**.

Our Model: Different from the above classical NLDP model where only one private dataset $\{(x_i, y_i)\}_{i=1}^n$ exists, the NLDP model in our setting allows the server to have an additional public

but unlabeled dataset $D' = \{x_j\}_{j=n+1}^{n+m} \subset \mathcal{X}^m$, where each x_j is sampled from $\mathcal{P}_x$, which is the marginal distribution of $\mathcal{P}$ (*i.e.,* they have the same distribution as $\{x_i\}_{i=1}^n$).

3 Privately Learning Generalized Linear Models

In this section, we study GLM in our model and privately estimate w^* in (1) by using both the private data $\{(x_i, y_i)\}_{i=1}^n$ and the public unlabeled data $\{x_j\}_{j=n+1}^{n+m}$. Our goal is to achieve a fully polynomial sample complexity for n and m, *i.e.,* $n, m = \text{Poly}(p, \frac{1}{\epsilon}, \frac{1}{\alpha}, \log\frac{1}{\delta})$, such that there is an (ϵ, δ)-NLDP algorithm with estimation error less than α (with high probability). Before presenting our ideas, we first consider the following lemma for $x \sim \mathcal{N}(0, \Sigma)$, which is from Stein's lemma [4].

Algorithm 1 Non-interactive LDP for smooth GLM with public data

Input: Private data $\{(x_i, y_i)\}_{i=1}^n \subset (\mathbb{R}^p \times \{0,1\})^n$, where $\|x_i\|_1 \le r$ and $|y_i| \le 1$, public unlabeled data $\{x_j\}_{j=n+1}^{n+m}$, loss function $\Phi : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters ϵ, δ, and initial value $c \in \mathbb{R}$.

1: **for** Each user $i \in [n]$ **do**
2: Release $\widehat{x_i x_i^T} = x_i x_i^T + E_{1,i}$, where $E_{1,i} \in \mathbb{R}^{p\times p}$ is a symmetric matrix and each entry of the upper triangle matrix is sampled from $\mathcal{N}(0, \frac{32r^4 \log\frac{2.5}{\delta}}{\epsilon^2})$.
3: Release $\widehat{x_i y_i} = x_i y_i + E_{2,i}$, where $E_{2,i} \in \mathbb{R}^p$ is sampled from $\mathcal{N}(0, \frac{32r^2 \log\frac{2.5}{\delta}}{\epsilon^2} I_p)$.
4: **end for**
5: **for** The server **do**
6: Let $\widehat{X^T X} = \sum_{i=1}^n \widehat{x_i x_i^T}$ and $\widehat{X^T y} = \sum_{i=1}^n \widehat{x_i y_i}$. Calculate $\hat{w}^{ols} = (\widehat{X^T X})^{-1}\widehat{X^T y}$.
7: Calculate $\tilde{y}_j = x_j^T \hat{w}^{ols}$ for each $j = n+1, \cdots, n+m$. Find the root $\hat{c}_\Phi$ such that $1 = \frac{\hat{c}_\Phi}{m}\sum_{j=n+1}^{n+m} \Phi^{(2)}(\hat{c}_\Phi \tilde{y}_j)$ using Newton's root-finding method:
8: **for** $t = 1, 2, \cdots$ until convergence **do**
9: $c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m+1} \Phi^{(2)}(c\tilde{y}_j) - 1}{\frac{1}{m}\sum_{j=n+1}^{n+m+1}\{\Phi^{(2)}(c\tilde{y}_j) + c\tilde{y}_j\Phi^{(3)}(c\tilde{y}_j)\}}$.
10: **end for**
11: **end for**
12: **return** $\hat{w}^{glm} = \hat{c}_\Phi \cdot \hat{w}^{ols}$.

Lemma 1 ([4])**.** If $x \sim \mathcal{N}(0, \Sigma)$, then w^* in (1) can be written as $w^* = c_\Phi \times w^{ols}$, where c_Φ is the fixed point of $z \mapsto (\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)])^{-1}$ (assuming $\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)] \neq 0$) and $w^{ols} = \Sigma^{-1}\mathbb{E}[xy]$ is the Ordinary Least Squares (OLS) vector.

From Lemma 1, we can see that to obtain w^*, it is sufficient to estimate w^{ols} and the underlying constant c_Φ. Specifically, to estimate w^{ols} in a non-interactive local differentially private way, a direct way is to let each player perturb her sufficient statistics, *i.e.,* $x_i x_i^T$ and $y_i x_i$. After receiving the private OLS estimator $\hat{w}^{ols}$, the server can then estimate the constant c_Φ by using the public unlabeled data and $\hat{w}^{ols}$. From the definition, it is easy to see that c_Φ is independent of the label y. Thus, c_Φ can be estimated by using the empirical version of $\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)]$. That is, find the root of the function $1 - \frac{c}{m}\sum_{j=n+1}^{n+m} \Phi^{(2)}(c\langle x_j, \hat{w}^{ols}\rangle)$. Several methods are available for finding roots, such as the Newton's method which has a quadratic convergence rate.

One problem with the above approach is that Lemma 1 needs x to be Gaussian, which implies that the sensitivity of the term $x_i x_i^T$ could be unbounded. We also note that Lemma 1 is only for Gaussian distribution. The following lemma extends Lemma 1 to bounded sub-Gaussian with an additional additive error of $O(\frac{\|w^*\|_\infty^2}{\sqrt{p}})$.

Lemma 2 ([10])**.** Let $x_1, \cdots, x_n \in \mathbb{R}^p$ be i.i.d realizations of a random vector x that is sub-Gaussian with zero mean, whose covariance matrix Σ has its corresponding $\Sigma^{\frac{1}{2}}$ being diagonally dominant [2], and whose distribution is supported on a ℓ_2-norm ball of radius r. Let $v = \Sigma^{-\frac{1}{2}}x$ be the whitened random vector of x with sub-Gaussian norm $\|v\|_{\psi_2} = \kappa_x$. If each v_i has constant

[2] A square matrix is said to be diagonally dominant if, for every row of the matrix, the magnitude of the diagonal entry in a row is larger than or equal to the sum of the magnitudes of all the other (non-diagonal) entries in that row.

first and second conditional moments (*i.e.,* $\forall j \in [p]$ and $\tilde{w} = \Sigma^{\frac{1}{2}} w^*$, $\mathbb{E}[v_{ij}|\sum_{k\neq j} \tilde{w} v_{ik}]$ and $\mathbb{E}[v_{ij}^2|\sum_{k\neq j} \tilde{w} v_{ik}]$ are deterministic) and the function $\Phi^{(2)}$ is Lipschitz continuous with constant G, then for $c_\Phi = \frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i, w^*\rangle)]}$ (assuming $\mathbb{E}[\Phi^{(2)}(\langle x_i, w^*\rangle)] \neq 0$), the following holds for GLM in (1)

$$\|\frac{1}{c_\Phi} \cdot w^* - w^{ols}\|_\infty \leq 16 G r \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}, \tag{2}$$

where ρ_q for $q = \{2, \infty\}$ is the conditional number of Σ in ℓ_q norm and $w^{ols} = (\mathbb{E}[xx^T])^{-1}\mathbb{E}[xy]$ is the OLS vector.

Lemma 2 indicates that we can use the same idea as above to estimate w^*. Note that the forms of c_Φ in Lemmas 1 and 2 are different. However, due to the closeness of w^* and w^{ols} in (2), we can still use $\frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i, w^{ols}\rangle \bar{c}_\Phi])}$ to approximate c_Φ, where $\bar{c}_\Phi$ is the root of $c\mathbb{E}[\Phi^{(2)}(\langle x_i, w^{ols}\rangle c) - 1$ (see Appendix for the details of the proof). Combining these ideas, we have Algorithm 1.

Theorem 1. For any $0 < \epsilon, \delta < 1$, Algorithm 1 is (ϵ, δ) non-interactive LDP.

The following theorem shows the sample complexity of the bounded sub-Gaussian case.

Theorem 2. Under the assumptions of Lemma 2, if further assume that the distribution of x is supported on the ℓ_1-norm ball with radius r, $|\Phi^{(2)}(\cdot)| \leq L$, and for some constant $\bar{c}$ and $\tau > 0$, the function $f(c) = c\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle c)]$ satisfies the condition of $f(\bar{c}) \geq 1 + \tau$, and the derivative of f in the interval $[0, \max\{\bar{c}, c_\Phi\}]$ does not change the sign (*i.e.,* its absolute value is lower bounded by some constant $M > 0$), then for sufficiently large m, n such that

$$m \geq \Omega\big(\|\Sigma\|_2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\rho_2\rho_\infty^2 p^2 \max\{1, \frac{1}{c_\Phi}\}^2\big) \tag{3}$$

$$n \geq \Omega\big(\frac{\rho_2\rho_\infty^2\|\Sigma\|_2^2 p^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \log\frac{1}{\delta} \log\frac{p}{\xi}}{\epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2\big), \tag{4}$$

with probability at least $1 - \exp(-\Omega(p)) - \xi$, the output $\hat{w}^{glm}$ in Algorithm 1 satisfies

$$\begin{aligned}\|\hat{w}^{glm} - w^*\|_\infty \leq O\big(&\frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}} p}{\sqrt{m}} \times \max\{\frac{1}{c_\Phi}, 1\}^2 \\ &+ \frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}} p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} \max\{\frac{1}{c_\Phi}, 1\}^2 \\ &+ \rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_\infty \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\{1, \frac{1}{c_\Phi}\}\big),\end{aligned} \tag{5}$$

where $G, L, \tau, M, \bar{c}, r, \kappa_x$ are assumed to be $O(1)$ and thus omitted in the Big-O notations (see Appendix for the explicit form of m and n).

Theorem 2 suggests that if we omit all the other terms and assume that $\|w^*\|_\infty = O(1)$, then for any given error $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$, there is an (ϵ, δ)-LDP algorithm whose sample complexities of both private (n) and public unlabeled (m) data, to achieve an estimation error of α (in ℓ_∞-norm), are $\tilde{O}(p^2\epsilon^{-2}\alpha^{-2})$.

Note that there are some previous work on LDP linear regression. [18] proposed an algorithm with a sample complexity of $\tilde{O}(p\alpha^{-2}\epsilon^{-2})$ and [26] achieved a sample complexity of $O(\log p\alpha^{-4}\epsilon^{-2})$. It seems that our sample complexity for the more general GLM is worse than theirs. However, these results are not really comparable due to their different settings. Firstly, [18, 26] considered the optimization error and [25] measured the ℓ_2-norm statistical error, while we estimate the ℓ_∞-norm statistical error. Secondly, w^* is assumed to be bounded in ℓ_2-norm in [18], ℓ_1-norm in [26], and ℓ_∞-norm in ours. There is also a result on NLDP linear regression [25]. It relies on assumptions that $\|x\|_2 = O(\sqrt{p})$ and w^* is 1-sparse, which are not needed in ours.

In the Appendix section, we provide the proof and experimental study of our algorithm.

References

[1] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[2] Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. Model-agnostic private learning. In *Advances in Neural Information Processing Systems*, pages 7102–7112, 2018.

[3] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

[4] David R Brillinger. A generalized linear model with "gaussian" regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012.

[5] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

[6] Paramveer Dhillon, Yichao Lu, Dean P Foster, and Lyle Ungar. New subsampling algorithms for fast least squares regression. In *Advances in neural information processing systems*, pages 360–368, 2013.

[7] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.

[8] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[10] Murat A Erdogdu, Mohsen Bayati, and Lee H Dicker. Scalable approximations for generalized linear problems. *The Journal of Machine Learning Research*, 20(1):231–275, 2019.

[11] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[12] Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *International Conference on Machine Learning*, pages 555–563, 2016.

[13] Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484, 2014.

[14] Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*, pages 488–497, 2016.

[15] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[16] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[17] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[18] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.

[19] G. W. Stewart. Matrix perturbation theory, 1990.

[20] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and XiaoFeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *CoRR*, abs/1709.02753, 2017.

[21] Terence Tao. Topics in random matrix theory. *Graduate Studies in Mathematics*, 132, 2011.

[22] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[23] Di Wang, Marco Gaboardi, and Jinhui Xu. Empirical risk minimization in non-interactive local differential privacy revisited. In *Advances in Neural Information Processing Systems*, pages 965–974, 2018.

[24] Di Wang, Adam Smith, and Jinhui Xu. Noninteractive locally private learning of linear models via polynomial approximations. In *Algorithmic Learning Theory*, pages 897–902, 2019.

[25] Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, June 9-15, 2019*, 2019.

[26] Kai Zheng, Wenlong Mou, and Liwei Wang. Collect at once, use effectively: Making non-interactive locally private learning possible. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4130–4139. JMLR. org, 2017.

A Experiments

A.1 Evaluation on synthetic data

Experimental Setting For GLM we consider the problem of binary logistic loss *i.e.,* $\Phi(\langle x, w\rangle) = \ln\left(1+\exp\left(\langle x, w\rangle\right)\right)$ in (1). For the problem we first compare the squared relative error $\frac{\|\hat{w}-w^*\|_\infty^2}{\|w^*\|_\infty^2}$ with respect to different privacy parameters $\epsilon \in \{10, 5, 3, 2\}$. In these experiments, we estimate the squared relative error with the fixed dimensionality $p = 10$ and the population parameter $w^* = (1, 1, ..., 1)/\sqrt{p}$. The sample size n is chosen from the set $10^4 \cdot \{1, 3, 5, ..., 29\}$. We assume that the same amount of public unlabeled data are available. The features are generated independently from a Bernoulli distribution $\Pr\left(x_{i,j} = \pm\frac{1}{p}\right) = 0.5$ and the label is generated according to the logistic model. The results are shown in Figure 1a. We then evaluate the impact of the dimensionality. In these experiments, we fix the privacy parameters [3] $\epsilon = 10$ and tune the dimensionality $p \in \{5, 10, 12, 15\}$. w^*s are the same as above. The sample size takes values from $n \in 10^4 \cdot \{10, 12, 14, ..., 48\}$ and the same amount of public unlabeled data are assumed. The labels are generated as the same as above. Here we measure the performance directly by the relative error. For each experiments above, we run 1000 times and take the average of the errors. The results are shown in Figure 1b.

From Figure 1a, we can see that the square of relative error is inversely proportional to the number of samples n. In other words, in order to achieve relative error α, we only need the number of private samples $n \sim \frac{1}{\alpha^2}$ if we omit the dependency on the other parameters. Besides, we also observe that the square of relative error is proportional to $\frac{1}{\epsilon^2}$, which matches our theoretical result. From Figure 1b, we can see that the relative error increases as the dimensionality increases. It may seem a little weird that it is not linear with the dimensionality. We note that as the dimensionality p changes, some other parameters, for example, the l_2 norm of the covariance matrix and w^*_∞ also change, which bring other effects to the relative error.

A.2 Evaluation on real data

We conduct experiment for GLM with logistic loss on the Covertype dataset [8]. Before running our algorithm, we first normalize the data and remove some co-related features. After the pre-processing, the dataset contains 581012 samples and 44 features. There are seven possible values for the label. Since multinomial logistic regression can not be regarded as a Generalized Linear Model,

[3] Note that in the studies on LDP ERM, ϵ is always chosen as a large value such as [3].

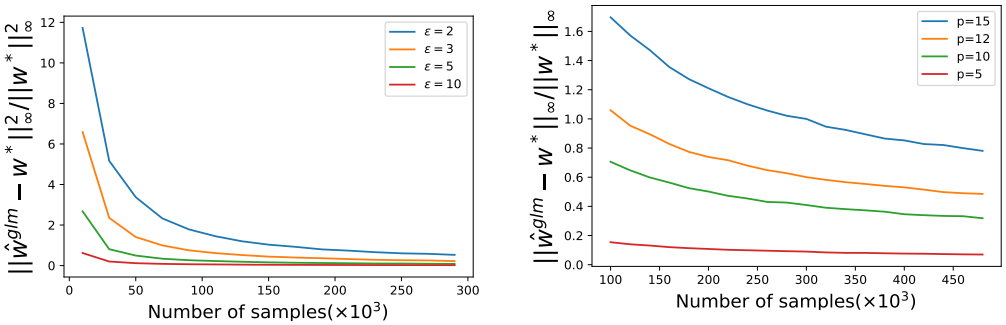

Figure 1: GLM with logistic loss under i.i.d Bernoulli design. The left plot shows the squared relative error under different levels of privacy. The right one shows relative error under different dimensionality.

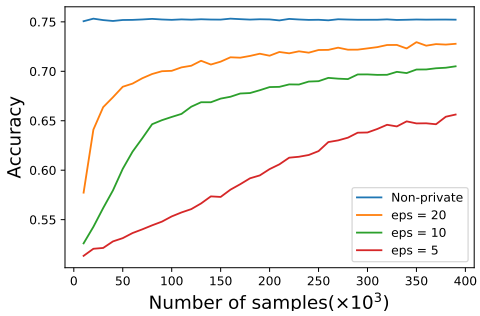

Figure 2: GLM with logistic loss on Covertype dataset

we consider a weaker test, which is to classify whether the label is Lodgepole Pine (type 2) or not. The algorithm that we choose is still binary logistic regression. We divide the data into training and testing, where $n_{\text{training}} = 406708$ and $n_{\text{testing}} = 174304$ and randomly choose the sample size $n \in 10^4 \cdot \{1, 2, 3, ..., 39\}$ from the training data and use exactly the same data as public. Regarding the privacy parameter, we let ϵ take value from $\{20, 10, 5\}$. We measure the performance by the prediction accuracy. For each combination of ϵ and n, the experiment is repeated 1000 times. We observe that when ϵ takes a reasonable value, the performance is approaching to the non-private case, provided that the size of private dataset is large enough. Thus, our algorithm is practical and is comparable to the non-private one.

B Background and Auxiliary Lemmas

Notations For a positive semi-definite matrix $M \in \mathbb{R}^{p\times p}$, we define the M-norm for a vector w as $\|w\|_M^2 = w^T M w$. $\lambda_{\min}(A)$ is the minimal singular value of the matrix A. For a semi positive definite matrix $M \in \mathbb{R}^{p\times p}$, let its SVD composition be $\Sigma = U^T \Sigma U$, where $\Sigma = \text{diag}(\lambda_1, \cdots, \lambda_p)$, then $M^{\frac{1}{2}}$ is defined as $M^{\frac{1}{2}} = U^T \Sigma^{\frac{1}{2}} U$, where $\Sigma^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_p})$.

Definition 2 (Sub-Gaussian)**.** For a given constant κ, a random variable $x \in \mathbb{R}$ is said to be sub-Gaussian if it satisfies $\sup_{m\geq 1} \frac{1}{\sqrt{m}} \mathbb{E}[|x|^m]^{\frac{1}{m}} \leq \kappa$. The smallest such κ is the **sub-Gaussian norm** of x and it is denoted by $\|x\|_{\psi_2}$. A random vector $x \in \mathbb{R}^p$ is called a sub-Gaussian vector if there exists a constant κ such that for any unit vector v, we have $\|\langle x, v\rangle\|_{\psi_2} \leq \kappa$.

Lemma 3 (Weyl's Inequality [19])**.** Let $X, Y \in \mathbb{R}^{p\times p}$ be two symmetric matrices, and $E = X - Y$. Then, for all $i = 1, \cdots, p$, we have

$$|\sigma_i(X) - \sigma_i(Y)| \leq \|E\|_2.$$

Lemma 4. Let $w \in \mathbb{R}^p$ be a fixed vector and E be a symmetric Gaussian random matrix where the upper triangle entries are i.i.d Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then, with probability at least $1 - \xi$, the following holds for a fixed positive semi-definite matrix $M \in \mathbb{R}^{p \times p}$

$$\|Ew\|_M^2 \leq \sigma^2 \text{Tr}(M) \|w\|^2 \log \frac{2p^2}{\xi}.$$

Proof of Lemma 4. Let $M = U^T \Sigma U$ denote the eigenvalue decomposition of M. Then, we have

$$\|Ew\|_M^2 = w^T E^T U^T \Sigma U E w = \sum_{i=1}^{p} \sigma_i \sum_{j=1}^{p} [UE]_{ij}^2 w_i^2.$$

Note that $[UE]_{i,j} = \sum_{k=1}^{p} U_{i,k} E_{j,k}$ where $E_{i,j}$ is Gaussian. Since U is orthogonal, we know that $[UE]_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Using the Gaussian tail bound for all $i, j \in [d]^2$, we have

$$\mathbb{P}(\max_{i,j \in [p]^2} |[UE]_{i,j} \geq \sqrt{\sigma^2 \log \frac{2p^2}{\xi}}) \leq \xi.$$

□

Lemma 5 (Theorem 4.7.1 in [22])**.** Let x be a random vector in $\mathbb{R}^p$ that is sub-Gaussian with covariance matrix Σ and $\|\Sigma^{-\frac{1}{2}} x\|_{\psi_2} \leq \kappa_x$. Then, with probability at least $1 - \exp(-p)$, the empirical covariance matrix $\frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$ satisfies

$$\|\frac{1}{n} X^T X - \Sigma\|_2 \leq C \kappa_x^2 \sqrt{\frac{p}{n}} \|\Sigma\|_2.$$

Lemma 6 (Corollary 2.3.6 in [21])**.** Let $M \in \mathbb{R}^{p \times p}$ be a symmetric matrix whose entries m_{ij} are independent for $j > i$, have mean zero, and are uniformly bounded in magnitude by 1. Then, there exists absolute constants $C_2, c_1 > 0$ such that with probability at least $1 - \exp(-C_2 c, 1p)$, the following inequality holds $\|M\|_2 \leq C\sqrt{p}$.

Below we introduce some concentration lemmas given in [10].

Lemma 7. Let $\mathbb{B}^\delta(\tilde{w})$ denote the ball centered at $\tilde{w}$ and with radius δ (*i.e.,* $\mathbb{B}^\delta(\tilde{w}) = \{w : \|w - \tilde{w}\|_2 \leq \delta\}$). For $i = 1, 2 \cdots, n$, let $x_i \in \mathbb{R}^p$ be i.i.d isotropic sub-Gaussian random vectors with $\|x_i\|_{\psi_2} \leq k_x$, and $\tilde{\mu} = \frac{\mathbb{E}[\|x\|_2]}{\sqrt{p}}$. For any given function $g : \mathbb{R} \mapsto \mathbb{R}$ that is Lipschitz continuous with G and satisfies $\sup_{w \in \mathbb{B}^\delta(\tilde{w})} \|g(\langle x, w \rangle)\|_{\psi_2} \leq \kappa_g$, with probability at least $1 - 2\exp(-p)$, the following holds for $np > 51 \max\{\chi, \chi^2\}$

$$\sup_{w \in \mathbb{B}^\delta(\tilde{w})} |\frac{1}{m} \sum_{i=1}^{m} g(\langle x_i, w \rangle) - \mathbb{E}[g(\langle x, w \rangle)]| \leq c(\kappa_g + \frac{\kappa_x}{\tilde{u}}) \sqrt{\frac{p \log m}{m}},$$

where $\chi = \frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{c\delta^2 G^2 \tilde{\mu}^2}$. c is some absolute constant.

Lemma 8. Let $\mathbb{B}^\delta(\tilde{w})$ be the ball centered at $\tilde{w}$ and with radius δ (*i.e.,* $\mathbb{B}^\delta(\tilde{w}) = \{w : \|w - \tilde{w}\|_2 \leq \delta\}$). For $i = 1, 2 \cdots, n$, let $x_i \in \mathbb{R}^p$ be i.i.d sub-Gaussian random vectors with covariance matrix Σ. For any given function $g : \mathbb{R} \mapsto \mathbb{R}$ that is uniformly bounded by L and Lipschitz continuous with G, the following holds with probability at least $1 - \exp(-p)$

$$\sup_{w \in \mathbb{B}^\delta(\tilde{w})} |\frac{1}{m} \sum_{i=1}^{m} g(\langle x_i, w \rangle) - \mathbb{E}[g(\langle x, w \rangle)]| \leq 2\{G(\|\tilde{w}\|_2 + \delta)\|\Sigma\|_2 + L\} \sqrt{\frac{p}{m}}.$$

The following lemma shows that the private estimator $\hat{w}^{ols}$ is close to the unperturbed one.

Lemma 9. Let $X = [x_1^T; x_2^T; \cdots; x_n^T] \in \mathbb{R}^{n \times d}$ be a matrix such that $X^T X$ is invertible, and $x_1, \cdots, x_n$ are realizations of a sub-Gaussian random variable x which satisfies the condition of $\|\Sigma^{-\frac{1}{2}} x\|_{\psi_2} \leq \kappa_x = O(1)$ and $\Sigma = \mathbb{E}[x x^T]$ is the the population covariance matrix. Let

$\tilde{w}^{ols} = (X^T X)^{-1} X^T y$ denote the empirical linear regression estimator. Then, for sufficiently large $n \geq \Omega(\frac{\kappa_x^4 \|\Sigma\|_2^2 p r^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)})$, the following holds with probability at least $1 - \exp(-\Omega(p)) - \xi$,

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\big(\frac{p r^2 (1 + r^2 \|\tilde{w}^{ols}\|_2^2) \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2 n \lambda_{\min}^2(\Sigma)}\big), \tag{6}$$

where $r = r$ if x_i is sampled from some bounded distribution.

Proof of Lemma 9. It is obvious that $\widehat{X^T X} = X^T X + E_1$, where E_1 is a symmetric Gaussian matrix with each entry sampled from $\mathcal{N}(0, \sigma_1^2)$ and $\sigma_1^2 = O(\frac{n r^4 \log \frac{1}{\delta}}{\epsilon^2})$. $\widehat{X^T y} = X^T y + E_2$, where E_2 is a Gaussian vector sampled from $\mathcal{N}(0, \sigma_2^2 I_p)$ and $\sigma_2^2 = O(\frac{n r^2 \log \frac{1}{\delta}}{\epsilon^2})$.

We first show that $\widehat{X^T X}$ is invertible with high probability under our assumption.

It is sufficient to show that $X^T X + E_1 \succ \frac{X^T X}{2}$, *i.e.,* $\|E_1\|_2 \leq \frac{\lambda_{\min}(X^T X)}{2}$. By Lemma 6, we can see that with probability $1 - \exp(-\Omega(p))$,

$$\|E_1\|_2 \leq O(\frac{r^2 \sqrt{pn \log \frac{1}{\delta}}}{\epsilon}).$$

Also, by Lemma 5 and Lemma 3 we know that with probability at least $1 - \exp(-\Omega(p))$,

$$\lambda_{\min}(X^T X) \geq n \lambda_{\min}(\Sigma) - O(\kappa_x^2 \|\Sigma\|_2 \sqrt{pn}).$$

Thus, it is sufficient to show that $n\lambda_{\min}(\Sigma) \geq O(\frac{\kappa_x^2 \|\Sigma\|_2 r^2 \sqrt{pn \log \frac{1}{\delta}}}{\epsilon})$, which is true under the assumption of $n \geq \Omega(\frac{\kappa_x^4 \|\Sigma\|_2^2 p r^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)})$. Thus, with probability at least $1 - \exp(-\Omega(p))$, it is invertible. In the following we will always assume that this event holds.

By direct calculation we have

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2 = -(X^T X + E_1)^{-1} E_1 \tilde{w}^{ols} + (X^T X + E_1)^{-1} E_2.$$

Thus, by Cauchy-Schwartz inequality we get

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\big(\|E_1 \tilde{w}^{ols}\|_{(X^T X + E_1)^{-2}}^2 + \|E_2\|_{(X^T X + E_1)^{-2}}^2\big).$$

Since we already assume that $X^T X + E_1 \succ \frac{X^T X}{2}$, by Lemma 4 we can obtain the following with probability at least $1 - \xi$

$$\|E_1 \tilde{w}^{ols}\|_{(X^T X + E_1)^{-2}}^2 \leq O\big(\frac{n r^4 \log \frac{1}{\delta}}{\epsilon^2} \|\tilde{w}^{ols}\|_2^2 \mathrm{Tr}((X^T X)^{-2}) \log \frac{4p^2}{\xi}\big)$$

$$\|E_2\|_{(X^T X + E_1)^{-2}}^2 \leq O\big(\frac{n r^2 \log \frac{1}{\delta}}{\epsilon^2} \mathrm{Tr}((X^T X)^{-2}) \frac{4p}{\xi}\big).$$

Thus, we have

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 \leq C_1 n \cdot \frac{r^2 (1 + r^2 \|\tilde{w}^{ols}\|_2^2) \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2} \mathrm{Tr}((X^T X)^{-2}).$$

For the term of $\mathrm{Tr}((X^T X)^{-2})$, we get

$$\mathrm{Tr}((X^T X)^{-2}) \leq (\mathrm{Tr}((X^T X)^{-1}))^2 \leq p \|(X^T X)^{-2}\|_2^2 = \frac{p}{\lambda_{\min}^2(X^T X)} \leq O(\frac{p}{n^2 \lambda_{\min}^2(\Sigma)}),$$

where the last inequality is due to the fact that $\lambda_{\min}(X^T X) \geq n\lambda_{\min}(\Sigma) - O(\kappa_x^2 \|\Sigma\|_2 \sqrt{pn}) \geq \frac{1}{2} n \lambda_{\min}(\Sigma)$ (by the assumption on n). This completes the proof. □

Let $w^{ols} = (\mathbb{E}[x x^T])^{-1} \mathbb{E}[xy]$ denote the population linear regression estimator. The following lemma bounds the estimation error between $\tilde{w}^{ols}$ and w^{ols}. The proof could be found in [10] or [6].

Lemma 10 (Prop. 7 in [10])**.** Assume that $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i x_i^T] = \Sigma$, and $\Sigma^{-\frac{1}{2}} x_i$ and y_i are sub-Gaussian with norms κ_x and γ, respectively. If $n \geq \Omega(\kappa_x \gamma p)$, the following holds

$$\|\tilde{w}^{ols} - w^{ols}\|_2 \leq O\big(\gamma \kappa_x \sqrt{\frac{p}{n \lambda_{\min}(\Sigma)}}\big),$$

with probability at least $1 - 3\exp(-p)$.

C Proofs

In order to show Theorem 2, we first show a theorem which is a generalization of theorem 2.

Theorem 3. Under the assumptions of Lemma 2, if further assume that the distribution of x is supported on the ℓ_1-norm ball with radius r, $\sup_{w:\|w-\Sigma^{\frac{1}{2}}w^{ols}\|_2\leq 1}\|\Phi^{(2)}(\langle x,w\rangle)\|_{\psi_2}\leq\kappa_g$, the function $f(c)=c\mathbb{E}[\Phi^{(2)}(\langle x,w^{ols}\rangle c)]$ satisfies the inequality of $f(\bar{c})\geq 1+\tau$ for some constant $\bar{c}$ and $\tau>0$, and the derivative of f in the interval of $[0,\max\{\bar{c},c_\Phi\}]$ does not change the sign (*i.e.,* its absolute value is lower bounded by some constant $M>0$), then for sufficiently large m,n such that

$$m\geq\tilde{\Omega}\big(\frac{1}{\tilde{\mu}^2}\epsilon^2 n\big), \tag{7}$$

$$n\geq\Omega\big(\|\Sigma\|_2^2\frac{p^2\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big), \tag{8}$$

the following holds with probability at least $1-\exp(-\Omega(p))-\xi$,

$$\begin{aligned}\|\hat{w}^{glm}-w^*\|_\infty\leq O\big(&\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\times\frac{p\|w^*\|_\infty\max\{1,\|w^*\|_\infty^3\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\\&+\rho_2\rho_\infty^2\frac{\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}+\\&\sqrt{\rho_2}\rho_\infty\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}\frac{1}{\tilde{\mu}}\sqrt{\frac{p^2\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}\big),\end{aligned} \tag{9}$$

where $\tilde{\mu}=\frac{\mathbb{E}[\|x\|_2]}{\sqrt{p}}$, the terms of $r,\kappa_x,\kappa_g,G,M,\tau,\bar{c}$ are assumed to be constants, and thus omitted in the Big-O notations (see Appendix for the explicit forms of m and n).

Since Theorem 3 is the most complicated one, we will first prove it and then Theorem 2.

C.1 Proof of Theorem 3

Since $r=O(1)$ (by assumption), combining this with Lemmas 9 and 10, we have that with probability at least $1-\exp(-\Omega(p))-\xi$ and under the assumption on n, there is a constant $C_3>0$ such that

$$\|\hat{w}^{ols}-w^{ols}\|_2\leq C_3\frac{\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}. \tag{10}$$

Lemma 11. Let $\Phi^{(2)}$ be a function that is Lipschitz continuous with constant G, and $f:\mathbb{R}\times\mathbb{R}^p\mapsto\mathbb{R}$ be another function such that $f(c,w)=c\mathbb{E}[\Phi^{(2)}(\langle x,w\rangle c)]$ and its empirical one is

$$\hat{f}(c,w)=\frac{c}{m}\sum_{j=1}^m\Phi^{(2)}(\langle x,w\rangle c).$$

Let $\mathbb{B}^\delta(\bar{w}^{ols})=\{w:\|w-\bar{w}^{ols}\|_2\leq\delta\}$, where $\bar{w}^{ols}=\Sigma^{\frac{1}{2}}w^{ols}$. Under the assumptions in Lemma 9 and Eq. (10), if further assume that $\|\Sigma^{-\frac{1}{2}}x\|_{\psi_2}\leq\kappa_x$, $\sup_{w\in\mathbb{B}^\delta(\bar{w}^{ols})}\|\Phi^{(2)}(\langle x,w\rangle)\|_{\psi_2}\leq\kappa_g$, and there exist $\bar{c}>0$ and $\tau>0$ such that $f(\bar{c},w^{ols})\geq 1+\tau$, then there is $\bar{c}_\Phi\in(0,\bar{c})$ such that $1=f(\bar{c}_\Phi,w^{ols})$. Also, for sufficiently large n and m such that

$$m\geq\Omega\big((\kappa_g+\frac{\kappa_x}{\tilde{\mu}})^2\max\{p\log m\tau^{-2},\frac{1}{G^2\tilde{\mu}^2}\frac{\epsilon^2 n}{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}\|\Sigma\|_2}\}\big), \tag{11}$$

$$n\geq\Omega(\kappa_x^4G^2\bar{c}^4\|\Sigma\|_2\frac{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big), \tag{12}$$

with probability at least $1-2\exp(-p)$, there exists a $\hat{c}_\Phi \in [0,\bar{c}]$ such that $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$. Furthermore, if the derivative of $c \mapsto f(c, w^{ols})$ is bounded below in the absolute value (*i.e.,* does not change sign) by $M > 0$ in the interval $c \in [0,\bar{c}]$, then the following holds

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\big(M^{-1}\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\big). \tag{13}$$

Proof of Lemma 11. We divide the proof into three parts.

Part 1: Existence of $\bar{c}_\Phi$: From the definition, we know that $f(0, w^{ols}) = 0$ and $f(\bar{c}, w^{ols}) > 1$. Since f is continuous, we known that there exists a constant $\bar{c}_\Phi \in (0,\bar{c})$ which satisfying $f(\bar{c}_\Phi, w^{ols}) = 0$.

Part 2: Existence of $\hat{c}_\Phi$: For simplicity, we use the following notations.

$$\delta = C_3\frac{\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}, \delta' = \frac{\|\Sigma\|_2^{\frac{1}{2}}\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}, \tag{14}$$

where C_3 is the one in (10). Thus, $\|\Sigma^{\frac{1}{2}}\hat{w}^{ols} - \Sigma^{\frac{1}{2}}w^{ols}\|_2 \leq \delta'$.

Now consider the term of $|\hat{f}(c,\hat{w}^{ols}) - f(c,\hat{w}^{ols})|$ for $c \in [0,\bar{c}]$. We have

$$\sup_{c\in[0,\bar{c}]}|\hat{f}(c,\hat{w}^{ols}) - f(c,\hat{w}^{ols})| \leq \sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})}|\hat{f}(c,w) - f(c,w)|, \tag{15}$$

where $\mathbb{B}_\Sigma^{\delta'}(w^{ols}) = \{w : \|\Sigma^{\frac{1}{2}}w - \Sigma^{\frac{1}{2}}w^{ols}\|_2 \leq \delta'\}$.

Note that for any x, we have $\langle x, w\rangle = \langle v, \Sigma^{\frac{1}{2}}w\rangle$, where $v = \Sigma^{-\frac{1}{2}}x$ follows an isotropic sub-Gaussian distribution. Also, by definition we know that $w \in \mathbb{B}_\Sigma^{\delta'}(w^{ols})$ is equivalent to $\Sigma^{\frac{1}{2}}w \in \mathbb{B}^{\delta'}(\bar{w}^{ols})$. Thus, we have

$$\begin{aligned}
&\sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})}|\hat{f}(c,\hat{w}^{ols}) - f(c,\hat{w}^{ols})| \\
&\leq \bar{c}\sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})}|\frac{1}{m}\sum_{j=1}^m\Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w\rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w\rangle c)| \\
&= \bar{c}\sup_{c\in[0,\bar{c}]}\sup_{\Sigma^{\frac{1}{2}}w\in\mathbb{B}^{\delta'}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m\Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w\rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w\rangle c)| \\
&= \bar{c}\sup_{w'\in\mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m\Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)|.
\end{aligned} \tag{16}$$

By Lemma 7, we know that when $mp \geq 51\max\{\chi,\chi^{-1}\}$, where

$$\chi = \frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{c\delta'^2G^2\tilde{\mu}^2} = \Theta\big(\frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{G^2\tilde{\mu}^2}\frac{\epsilon^2 n\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}\|\Sigma\|_2}\big),$$

the following holds with probability at least $1-2\exp(-p)$

$$\sup_{w'\in\mathbb{B}^{\bar{c}\delta}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m\Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)| \leq O((\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}). \tag{17}$$

By the Lipschitz property of $\Phi^{(2)}$, we have that for any w_1 and w_2,

$$\begin{aligned}
\sup_{c\in[0,\bar{c}]}|f(c,w_1) - f(c,w_2)| &\leq G\bar{c}^2\mathbb{E}[\langle v, \Sigma^{\frac{1}{2}}(w_1 - w_2)\rangle] \\
&\leq \kappa_x G\bar{c}^2\|\Sigma^{\frac{1}{2}}(w_1 - w_2)\|_2.
\end{aligned} \tag{18}$$

Taking $w_1 = \hat{w}^{ols}$ and $w_2 = w^{ols}$, we have

$$\sup_{c\in[0,\bar{c}]} |f(c,\hat{w}^{ols}) - f(c,w^{ols})| \le O\big(\kappa_x G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}\big).$$

Combining this with (16), (17), (18), and taking δ as in (14), we get

$$\sup_{c\in[0,\bar{c}]} |\hat{f}(c,\hat{w}^{ols}) - f(c,w^{ols})| \le O\big(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\big). \tag{19}$$

Let B denote the RHS of (19). If $c = \bar{c}$, we have $\hat{f}(c,\hat{w}^{ols}) \ge 1 + \tau - B$. Thus, if $B \le \tau$, there must exist a $\hat{c}_\Phi \in [0,\bar{c}]$ such that $\hat{f}(\hat{c}_\Phi,\hat{w}^{ols}) = 1$.

To ensure that $B \le \tau$ holds, it is sufficient to have

$$O(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}) \le \frac{\tau}{2}$$

and

$$O(G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}) \le \frac{\tau}{2}.$$

This means that

$$m \ge \Omega\big(\bar{c}^2(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2 p\log m\tau^{-2}\big),$$

$$n \ge \Omega(\kappa_x^4 G^2\bar{c}^4\|\Sigma\|_2 \frac{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}),$$

which are assumed in the lemma.

Part 3: Estimation Error: So far, we know that $\hat{f}(\hat{c}_\Phi,\hat{w}^{ols}) = f(\bar{c}_\Phi,w^{ols}) = 1$ with high probability. By (15), (16) and (17), we have

$$|1 - f(\hat{c}_\Phi,\hat{w}^{ols})| = |\hat{f}(\hat{c}_\Phi,\hat{w}^{ols}) - f(\hat{c}_\Phi,\hat{w}^{ols})| \le O(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}).$$

By the same argument for (19), we have

$$|f(\hat{c}_\Phi,\hat{w}^{ols}) - f(\hat{c}_\Phi,w^{ols})| \le G\kappa_x\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}.$$

Thus, using Taylor expansion on $f(c,w^{ols})$ around c_Φ and by the assumption of the bounded derivative of f, we have

$$\begin{aligned} M|\hat{c}_\Phi - \bar{c}_\Phi| &\le |f(\hat{c}_\Phi,w^{ols}) - f(\bar{c}_\Phi,w^{ols})| \\ &\le |f(\hat{c}_\Phi,w^{ols}) - f(\hat{c}_\Phi,\hat{w}^{ols})| + |f(\hat{c}_\Phi,\hat{w}^{ols}) - 1| \\ &\le O\big(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\big). \end{aligned}$$

□

Next, we prove our main theorem.

Proof of Theorem 3. By definition, we have

$$\begin{aligned} \|\hat{w}^{glm} - w^*\|_\infty &\le \|\hat{c}_\Phi\hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - w^*\|_\infty \\ &\le \|\hat{c}_\Phi\hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty + \|c_\Phi w^{ols} - w^*\|_\infty. \end{aligned} \tag{20}$$

We first bound the term of $|\bar{c}_\Phi - c_\Phi|$. Since $\bar{c}_\Phi \mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle \bar{c}_\Phi)] = 1$ and $c_\Phi \mathbb{E}[\Phi^{(2)}(\langle x, w^*\rangle)] = 1$ (by definition), we get

$$\begin{aligned}
|f(\bar{c}_\Phi, w^{ols}) - f(c_\Phi, w^{ols})| &= |c_\Phi \mathbb{E}[\Phi^{(2)}(\langle x, w^*\rangle)] - f(c_\Phi, w^{ols})| \\
&\le c_\Phi |\mathbb{E}[\Phi^{(2)}(\langle x, w^*\rangle) - \Phi^{(2)}(\langle x, w^{ols}\rangle c_\Phi)] \\
&\le c_\Phi G |\mathbb{E}[\langle x, (w^* - c_\Phi w^{ols})\rangle] \\
&\le c_\Phi G \|(w^* - c_\Phi w^{ols})\|_\infty \mathbb{E}\|x\|_1 \\
&\le c_\Phi G r \|c_\Phi w^{ols} - w^*\|_\infty,
\end{aligned}$$

where the last inequality is due to the assumption that $\|x\|_1 \le r$.

Thus, by the assumption of the bounded deviation of $f(c, w^{ols})$ on $[0, \max\{\bar{c}, c_\Phi\}]$, we have

$$M|\bar{c}_\Phi - c_\Phi| \le |f(\bar{c}_\Phi, w^{ols}) - f(c_\Phi, w^{ols})| \le c_\Phi G r \|c_\Phi w^{ols} - w^*\|_\infty.$$

By Lemma 2 in the context, we have

$$|\bar{c}_\Phi - c_\Phi| \le 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}. \tag{21}$$

Thus, the second term of (20) is bounded by

$$\begin{aligned}
&\|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty \le 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}} \|w^{ols}\|_\infty \\
&\le 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^3}{\sqrt{p}} \Big(\frac{1}{c_\Phi} + 16 G r \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}}\Big) \\
&= O\big(M^{-1} r^3 \kappa_x^6 G^3 \rho_2 \rho_\infty^2 \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\{1, c_\Phi\}\big),
\end{aligned} \tag{22}$$

where the last inequality is due to Lemma 2 in the context.

By Lemma 2 in the context, the third term of (20) is bounded by $16 c_\Phi G r \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}$.

For the first term of (20), by (10) and Lemma 11 we have

$$\begin{aligned}
&\|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty \le |\hat{c}_\Phi| \cdot \|\hat{w}^{ols} - w^{ols}\|_\infty + |\hat{c}_\Phi - \bar{c}_\Phi| \cdot \|w^{ols}\|_\infty \\
&\le O\Big(\bar{c} \frac{\kappa_x \sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \\
&+ \|w^{ols}\|_\infty \Big(M^{-1} \bar{c} (\kappa_g + \frac{\kappa_x}{\tilde{\mu}}) \sqrt{\frac{p \log m}{m}} + M^{-1} G \kappa_x^2 \bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{\sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\Big)\Big).
\end{aligned} \tag{23}$$

For the first term of (23), we have

$$\begin{aligned}
&\bar{c} \frac{\kappa_x \sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \le \bar{c} \frac{\kappa_x p r^2 \|w^{ols}\|_\infty \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \\
&\le \bar{c} \frac{\kappa_x p r^2 \|w^*\|_\infty \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \Big(\frac{1}{c_\Phi} + 16 G r \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}}\Big) \\
&= O\Big(\bar{c} \frac{p \kappa_x^4 \sqrt{\rho_2} \rho_\infty G r^3 \|w^*\|_\infty \max\{1, \|w^*\|_\infty\} \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}\Big).
\end{aligned} \tag{24}$$

For the second term of (23), we have

$$\begin{aligned}
&\|w^{ols}\|_\infty M^{-1}\bar{c}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\\
&\le \bar{c}\|w^*\|_\infty(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})\\
&\le O\big(Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}\big). \qquad (25)
\end{aligned}$$

For the third term of (23), we have

$$\begin{aligned}
&\|w^{ols}\|_\infty M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}})\\
&\le M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{pr^2\|w^*\|_\infty^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})^2\\
&\le O\big(M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\frac{pr^4\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big). \qquad (26)
\end{aligned}$$

Thus, the first term of (20) is bounded by (since $m\ge\Omega(n)$)

$$\begin{aligned}
&\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty\le O\big(\bar{c}\frac{p\kappa_x^4\sqrt{\rho_2}\rho_\infty Gr^3\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}\\
&+Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}+\\
&M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\frac{pr^4\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\\
&=O\big(M^{-1}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\\
&\times\frac{pr^4\|w^*\|_\infty\max\{1,\|w^*\|_\infty^3\}\sqrt{\log m\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big).
\end{aligned}$$

Putting all the bounds together, we have

$$\begin{aligned}
&\|\hat{w}^{glm}-w^*\|_\infty\le\tilde{O}\big(M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\\
&\times\frac{pr^4\|w^*\|_\infty\max\{1,\|w^*\|_\infty^3\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\\
&+M^{-1}r^3\kappa_x^6c_\Phi G^3\rho_2\rho_\infty^2\frac{\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}+\\
&Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}\big). \qquad (27)
\end{aligned}$$

Next, we bound the probability. We assume that Lemma 9, 10 and 11 hold with probability at least $1-\exp(-\Omega(p))-\rho$. They hold when

$$m\ge\Omega\big((\kappa_g+\frac{\kappa_x}{\tilde{\mu}})^2\max\{p\log m\tau^{-2},\frac{1}{G^2\tilde{\mu}^2}\frac{\epsilon^2 n}{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}\}\big), \qquad (28)$$

$$n\ge\Omega(\max\{\kappa_x^4G^2\bar{c}^4\|\Sigma\|_2\frac{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}},\frac{\kappa_x^4\|\Sigma\|_2^2pr^4\log\frac{1}{\delta}}{\epsilon^2\lambda_{\min}^2(\Sigma)}\}\big). \qquad (29)$$

Since $\|w^{ols}\|_2 \leq \sqrt{p}\|w^*\|_\infty(\frac{1}{c_\Phi} + 16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})$, it suffices for n

$$n \geq \Omega\big(G^4\bar{c}^4\|\Sigma\|_2^2\frac{p^2r^6\kappa_x^{10}\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big). \tag{30}$$

□

C.2 Proof of Theorem 2

Lemma 12. Let $\bar{c}_\Phi, \bar{c}, \tau, f, \hat{f}$ be defined the same as in Lemma 11. If further assume that $|\Phi^{(2)}(\cdot)| \leq L$ for some constant $L > 0$ and is Lipschitz continuous with constant G, then, under the assumptions in Lemma 9 and (10), with probability at least $1 - 4\exp(-p)$ there exists a constant $\hat{c}_\Phi \in [0, \bar{c}]$ such that $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$. Furthermore, if the derivative of $c \mapsto f(c, w^{ols})$ is bounded below in absolute value (*i.e.,* does not change the sign) by $M > 0$ in the interval $c \in [0, \bar{c}]$, then with probability at least $1 - 4\exp(-p)$, the following holds

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\big(\frac{M^{-1}GL\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}} + M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) \tag{31}$$

for sufficiently large m, n such that

$$n \geq \Omega\big(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big) \tag{32}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2p\tau^{-2}\big). \tag{33}$$

Proof of Lemma 12 . The main idea of this proof is almost the same as the one for Lemma 11. The only difference is that instead of using Lemma 7 to get (17), we use here Lemma 8 to obtain the following with probability at least $1 - \exp(-p)$

$$\begin{aligned}
&\sup_{w'\in\mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m\Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)| \\
&\leq O\big((G(\|\bar{w}^{ols}\|_2 + \bar{c}\delta')\|I\|_2 + L)\sqrt{\frac{p}{m}} \\
&\leq O\big((G\|\Sigma\|_2^{\frac{1}{2}}(\|w^{ols}\|_2 + \bar{c}\frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}) + L)\sqrt{\frac{p}{m}}\big).
\end{aligned} \tag{34}$$

Thus, by (16), (18) and (34), we have

$$\begin{aligned}
\sup_{c\in[0,\bar{c}]}|\hat{f}(c,\hat{w}^{ols}) - f(c,w^{ols})| \leq O\big(&G\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}} + \\
&\frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\sqrt{\frac{p}{mn}} + L\sqrt{\frac{p}{m}}\big).
\end{aligned} \tag{35}$$

Let D denote the RHS of (35), we have

$$\hat{f}(\bar{c},\hat{w}^{ols}) \geq 1 + \tau - D.$$

It is sufficient to show that $\tau > D$, which holds when

$$O(G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}) \leq \frac{\tau}{2}$$

and

$$O(\frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}L\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\sqrt{\frac{p}{mn}}) \leq \frac{\tau}{2}.$$

That is,

$$n \geq \Omega\big(\frac{G^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4 p r^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big) \tag{36}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2 p\tau^{-2}\big). \tag{37}$$

Then, there exists $\hat{c}_\Phi \in [0,\bar{c}]$ such that $\hat{f}(\hat{c}_\Phi,\hat{w}^{ols})=1$. We can easily get

$$\begin{aligned}
M|\hat{c}_\Phi-\bar{c}_\Phi| &\leq |f(\hat{c}_\Phi,w^{ols})-f(\bar{c}_\Phi,w^{ols})| \\
&\leq O\big(\frac{G\bar{c}^2\kappa_x^2 r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}} \\
&+ \frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\sqrt{\frac{p}{mn}} + LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) \qquad (38)\\
&\leq O\big(\frac{GL\bar{c}^2\kappa_x^2 r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}} + LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big). \qquad (39)
\end{aligned}$$

□

Proof of Theorem 2 . The proof is almost the same as the one for Theorem 3. By definition, we have

$$\begin{aligned}
\|\hat{w}^{glm}-w^*\|_\infty &\leq \|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols}-w^*\|_\infty \\
&\leq \|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols}-c_\Phi w^{ols}\|_\infty + \|c_\Phi w^{ols}-w^*\|_\infty. \qquad (40)
\end{aligned}$$

The second term of (40) is bounded by

$$\|\bar{c}_\Phi w^{ols}-c_\Phi w^{ols}\|_\infty \leq O\big(M^{-1}r^2\kappa_x^7 c_\Phi G^3\rho_2\rho_\infty^2\frac{\|w^*\|_\infty^3\max\{1,\|w^*\|_\infty\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}\big). \tag{41}$$

By Lemma 2 in the context, the third term of (40) is bounded by $16c_\Phi Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}}$. The first term is bounded by

$$\begin{aligned}
&\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty \leq \\
&O\big(\frac{M^{-1}G^3L\bar{c}^2\kappa_x^8 r^4\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}\times\max\{\frac{1}{c_\Phi},1\}^2 \\
&+\frac{M^{-1}G^3L\bar{c}^2\kappa_x^6 r^2\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p}{\sqrt{m}}\times\max\{\frac{1}{c_\Phi},1\}^2\big). \qquad (42)
\end{aligned}$$

Thus, in total we have

$$\begin{aligned}
\|\hat{w}^{glm}-w^*\|_\infty &\leq O\big(\frac{M^{-1}G^3L\bar{c}^2\kappa_x^6 r^2\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p}{\sqrt{m}}\times\max\{\frac{1}{c_\Phi},1\}^2 \\
&+\frac{G^3L\bar{c}^2\kappa_x^6 r^4\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}\max\{\frac{1}{c_\Phi},1\}^2 \\
&+M^{-1}r^2\kappa_x^7 c_\Phi G^3\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_\infty\frac{\|w^*\|_\infty^3\max\{1,\|w^*\|_\infty\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}\big). \qquad (43)
\end{aligned}$$

The probability of success is at least $1-\exp(-\Omega(p))-\xi$. The sample complexity should satisfy

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}G^2r^2\kappa_x^6\rho_2\rho_\infty^2p^2\tau^{-2}\max\{1,\frac{1}{c_\Phi}\}^2\big) \tag{44}$$

$$n \geq \Omega\big(\frac{\rho_2\rho_\infty^2G^4\tau^{-2}\bar{c}^4\|\Sigma\|_2^2\kappa_x^{10}p^2\|w^*\|_\infty^2r^6\max\{1,\|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^3}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big). \tag{45}$$

□